IEEE Transactions on Industrial Electronics, Volume 69, Number 7, July 2022, pp. 7405-7414
DOI:10.1109/TIE.2021.3099254

Learning spatio-temporal representations with a dual-stream 3D residual network for non-driving activity recognition

Lichao Yang, Xiaocai Shan, James Brighton, Chen Lv, *Senior Member, IEEE* and Yifan Zhao*, *Senior Member, IEEE*

***Abstract*—Accurate recognition of non-driving activity (NDA) is important for the design of intelligent Human Machine Interface to achieve a smooth and safe control transition in the conditionally automated driving vehicle. However, some characteristics of such activities like limited-extent movement and similar background pose a challenge to the existing 3D convolutional neural network (CNN) based action recognition methods. In this paper, we propose a dual-stream 3D residual network, named DS3D ResNet, to enhance the learning of spatio-temporal representation and improve the activity recognition performance. Specifically, a parallel 2-stream structure is introduced to focus on the learning of short-time spatial representation and small-region temporal representation. A 2-feed driver behaviour monitoring framework is further build to classify 4 types of NDAs and 2 types of driving behaviour based on the driver's head and hand movement. A novel NDA dataset has been constructed for the evaluation, where the proposed DS3D ResNet achieves 83.35% average accuracy, at least 5% above three selected state-of-the-art methods. Furthermore, this study investigates the spatio-temporal features learned in the hidden layer through the saliency map, which explains the superiority of the proposed model on the selected NDAs.**

***Index Terms*—action recognition, non-driving related task, automated driving**

I. Introduction

More and more level 3 automated driving vehicles will be on road in the coming years [1], and such vehicles allow drivers to take their hands and eyes off the road. However, according to SAE (J3016) Automation Levels, in level 3, drivers are still expected to take control of the vehicle if there is a request to intervene [2]. The driver's situation awareness in terms of driving environment and vehicle condition is reduced since they do not need to pay full attention to road and dashboard, which could bring a risk when the driver takes over the vehicle control without a right process in place. Therefore, it is of great importance to monitor the driver's behaviour during the level 3 automated driving and design the specific takeover request modality or Human Machine Interface (HMI) for different states to ensure a smooth and safe control transition [3].

There are two kinds of activities that the driver could engage inside the vehicle cabin, which are driving activities (DAs) and non-driving activities (NDAs). Similar to distraction and fatigue, the engagement of NDAs could reduce the driver's situation awareness. Normally, the methods of detecting NDAs engagement is based on the driver's attention [4]. Since the drivers always check the road or surrounding environment when they are conducting DAs, while during NDAs engagement, they pay more attention to the object they are engaging with. Moreover, different NDAs could lead to different impacts on the driver's take-over performance [5]–[7]. A refined classification of NDAs could help to design an intelligent take-over process to improve driving safety. During NDAs engagement, the driver's hand movement contains information about the interaction between the driver and the object, which can be used for further classification. Therefore, both visual attention and behaviour are necessary for the recognition of the driver's activity in the vehicle.

The recognition of the driver's NDAs has been widely researched in the last few years. With the rapid development of deep learning in activity recognition based on videos, computer vision-based methods become the focus for NDAs recognition [3], [8], [9]. The methods for action recognition using videos can be roughly divided into two categories: spatio-temporal attention mechanisms and 3D convolutional neural network (CNN). Both methods employ the CNN for spatial feature extraction due to its great learning capability in the spatial domain. The spatio-temporal attention mechanisms learn the temporal features by employing the sequence-based signal processing methods like Recurrent Neural Network, Long Short-Term Memory and transformer [10], [11]. 3D CNN extends the 2D spatial features into 3D features by adding a convolutional kernel in the temporal domain [12]–[15]. For the NDAs recognition, unlike the traditional activities in the action recognition dataset [16], [17], such as Tai Chi, Basketball, Diving, etc., which contains diverse spatial information in the background and large-scale body movement, NDAs are

Manuscript received .... This work was supported by Cranfield's EPSRC Impact Accrelate Account EP/R511511/1.

L. Yang, J. Brighton and Y. Zhao are with School of Aerospace, Transport and Manufacturing, Cranfield University, Bedfordshire MK43 0AL, UK.

X. Shan is with Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China.

C. Lv is with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

*the corresponding author: Y. Zhao (yifan.zhao@cranfield.ac.uk)

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works

constrained in the vehicle cabin. Normally, the movement that matters is the driver's hand. The hand movement is more complex in the temporal domain and the background is similar in the spatial domain, which poses a challenge to the existing 3D CNN models [12], [18] for activity recognition. Considering that, a proper design of the 3D CNN model could enhance the spatio-temporal representations of the activity with less 3D convolutional computation to achieve good recognition performance. Furthermore, the driver's head movement is also needed to be evaluated, since the driver visual attention is also a key factor to determine the NDAs engagement. In this paper, we propose a 2-feed 3D CNN based driver behaviour recognition system. This system focuses on both driver's head and hand movement to recognise whether the driver is engaging with an NDA or not, and further determine the type of NDA or DA. We design a dual-stream 3D residual network, named DS3D ResNet, to enhance the short-time spatial representation and small-region temporal representation learned on separate streams. A novel NDA dataset has been produced to evaluate the proposed model and other state-of-the-art models. This study also visualises the hidden layers of the proposed model to further verify and explain the semantic features that the model learned

II. Related Work

NDAs recognition: The methods of activity recognition can be roughly divided into 2 categories from the perspective of feature extraction, which are hand-crafted features based methods and deep learning based methods. The first kind of method classifies the activities based on some hand-crafted features like driver's gaze direction, hand movement and body pose. Martin et al. [19] extracted features of the driver's upper body pose and proposed a 3-stream recurrent neural network (RNN) system. This system evaluates the spatial relationship of body joints, the temporal skeleton movement and the context of the driver's surrounding to recognise the selected NDAs, including drinking, phone texting, calling, reading and eating. Furthermore, Xing et al. [20] combined the depth information inside the vehicle cabin with the features mentioned above and established a feedforward neural network (FFNN) to identify the activities. Yang et al. [4] proposed a dual-camera gaze estimation system and addressed the NDA recognition problem from the perspective of the driver's eye. With the development of CNN in the field of activity recognition [12], [13], [21]–[23],

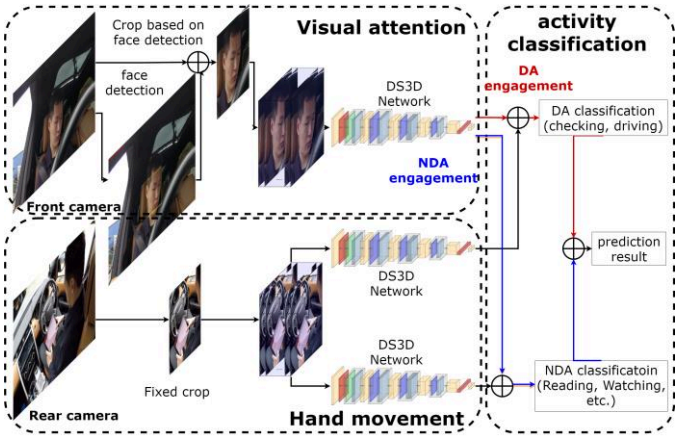


Fig. 1. Driver activity recognition framework.

the deep learning-based methods have attracted increasing attentions in NDA recognition in recent years. Xing et al. [8] removed the image background and used the drivers' body as the input of the CNN model to recognise their behaviours. Yang et al. [3] employed a 2-stream CNN model to extract the spatial features from the original image and the driver's hand movement features from the corresponding optical flow images. Moreover, Eraqi et al. [24] trained different CNNs on multiple inputs including raw images, skin-segmented images, face images, hands images, and "face+hands" images. The final prediction is obtained by using a genetic algorithm based on the outputs of all the CNN models.

3D CNN: CNN has been widely researched in recent years and made great achievement on the spatial representation, particularly in the scope of computer vision. CNN has been mainly applied to 2D images that lack the temporal representation, which is especially crucial for the application of video classification. To address this challenge, 3D CNN was employed to learn the spatio-temporal representations and extract the motion information hidden in the video frames [12], [25]. The residual structure [26] was implemented to tackle the training difficulty in the deeper 3D CNN model [18]. Since the computation cost of the deep 3D CNN is expensive and the model size is relatively large, Qiu et al. [27] proposed a Pseudo-3D network to factorise 3D convolutions into spatial convolutions and temporal convolutions to reduce the computational complexity. They compacted the model with 3 different forms of the spatio-temporal residual blocks. Similarly, Tran et al. [28] used only a spatial convolution followed by a temporal convolution residual block in the proposed R(2+1)D network and achieved better action recognition performance.

Unlike other deep learning-based methods for the NDA recognition, which mainly focus on the 2D image domain, our work attempt to extract the spatio-temporal features from the driver behaviour in the video domain. Considering the characterisation of NDAs, the capability of 3D CNN has not been fully exploited with the existing architecture mentioned above. In this work, we improve the spatial-temporal representation of residual blocks in the network with a designed dual-stream structure by enhancing the small-region temporal representation and the short-time spatial representation in different scales. The idea of this work is not only to revise the network structure but also to develop a framework to recognise and classify the type of NDA engagement during level 3 automated driving. The proposed framework is given in details in the next section.

III. Methodology

The proposed 2-feed dual-stream 3D residual network-based driver activity recognition framework is illustrated in Fig. 1. There are 2 feeds in this framework, which are the frames from the front camera and the rear camera. The front camera captures the driver's head movement and estimates the visual attention, which is used to recognise whether the driver is engaging with NDAs or not. The input of the 3D CNN model for this feed a stack of frames, which are cropped based on the location of the detected face from raw frames. The rear camera focuses on the driver's behaviour in the cabin mainly the hand movement,

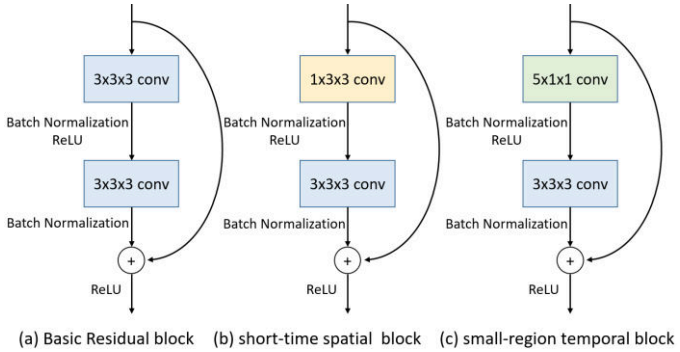


Fig.2. The basic residual block and the proposed blocks

which aims to further classify the specific NDAs or DAs. The final activity classification is obtained by combining these two results.

A. 3D Residual Block

3D convolution is the most natural method to extract the spatio-temporal features from videos [12], [27]. It has the capability to model the temporal connection among the spatial information encoded frames. For the 3D convolution, the filter is denoted as $d \times k \times k$, where d and k are the temporal depth and the spatial size of the filter respectively.

Following the success of the Residual Networks (ResNets) in encoding the spatio-temporal information for action recognition task [18], [26]. We propose 2 different residual blocks to enhance the short-time spatial representation and the small-region temporal representation of the model, as illustrated in Fig. 2(b) and 2(c), based on the basic residual block in Fig. 2(a). There are 2 convolutional layers in a basic residual block. Each layer is followed by batch normalization [29]. The filter size of each convolutional layer is $3 \times 3 \times 3$.

The output of the l-th residual block can be expressed as:

$$x_{l+1} = F(x_l, \{W_i\}) + x_l. \quad (1)$$

where x_{l+1} and x_l are the output and input of the block. The function $F(x_l, \{W_i\})$ is the learned residual mapping of the block and weight $\{W_i\}$ is for multiple convolutional layers.

The short-time spatial block (see Fig. 2(b)) aims to encode the change of the spatial information in a short time. Unlike the basic residual block, the size of the filter S used in the first convolutional layer of the proposed block is $1 \times 3 \times 3$. This filter compresses the temporal dimension, which is equivalent to the 2D convolutional filter on the spatial domain. The filter R of the second convolutional layer is still a $3 \times 3 \times 3$ filter to expand the receptive field in both temporal and spatial domains. The block can be expressed as:

$$x_{l+1} = R(S(x_l, W_s), W_r) + x_l. \quad (2)$$

The small-region temporal block, shown in Fig. 2(c), concentrates on a small area and captures its change in a long period. The size of the first convolutional filter (T) is $5 \times 1 \times 1$, which can be considered as a 1D convolutional filter on the temporal domain. It is followed by a $3 \times 3 \times 3$ filter (R). The output of this block can be expressed as:

$$x_{l+1} = R(T(x_l, W_t), W_r) + x_l. \quad (3)$$

The ReLU activation function is employed after the first convolutional layer and the output of all these blocks.

B. Architecture of the 3D CNN Model

The architecture of the network is illustrated in Fig. 3. For simplicity, the size of the given video clip is denoted as $c \times l \times h \times w$, where c is the number of channels, l is the number of frames in the clip, h and w are the height and width of images, respectively. The input of the network is a $3 \times 16 \times 112 \times 112$ tensor. The parallel structure is employed after the first convolution block. The upper spatial stream uses a sequence of 4 spatial blocks to emphasise the short-time spatial information in different scales. The bottom temporal stream has 4 temporal blocks connected in series, which focus on the change in the small-region temporal domain. After pooling, the size of the feature map for each stream is $256 \times 1 \times 1 \times 1$. The final 512-dimensional vector is obtained by concatenating the feature maps produced in both 2 streams and fed into a fully connected layer, which outputs the final prediction probabilities through the softmax function.

C. Prediction Process for the Framework

As illustrated in Fig. 1, the prediction of the driver activity recognition framework combines the outputs from 3 separate models. The prediction probability of NDA engagement recognition based on the driver's head movement is denoted as P_e, which has two states: DA engagement and NDA engagement, denoted as c_D and c_N, respectively. The prediction probability for these two classes is represented by $P_e[c_D]$ and $P_e[c_N]$. Two different 3D CNN models have been trained separately for NDA and NDA classification based on hand movement. The prediction probability for these 2 models are denoted as P_{Dc} and P_{Nc}. The final scores of the DA classification and NDA classification are denoted as Y_{d} and Y_{N}.

The score of a single DA can be expressed as:

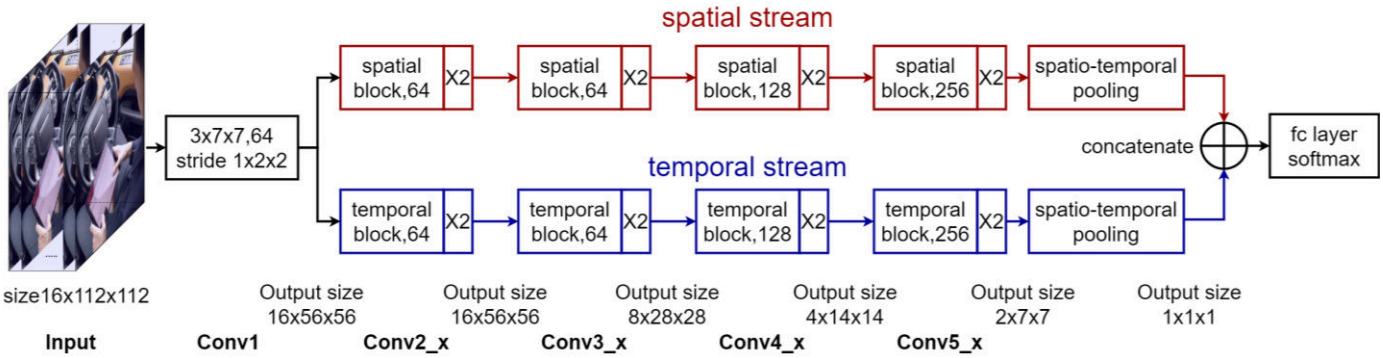


Fig. 3. The proposed network architecture. The layer name is the bolded word at the bottom. The output size of each layer is on the top right of the layer name. The details of the each used blocks is introduced in Fig. 2. Downsampling is employed on conv3_1, conv4_1, conv5_1 with a stride of 2.

$$Y_D(i_{nN}) = P_{Dc}(i_D)P_e[c_D], \quad (4)$$

where i_D is the index of the DAs. The score of a single NDA can be expressed as:

$$Y_N(i_N) = P_{Nc}(i_N)P_e[c_N], \quad (5)$$

where i_N is the index of the NDAs. The final prediction scores for all NDA and DA classes, denoted by Y, can be expressed as:

$$Y = Y_{\mathrm{D}} \cup Y_{\mathrm{N}}. \quad (6)$$

D. Visual Explanations of CNN Model Predictions

With the effort of visual explanation for CNN [30]–[32], we can explain the prediction of the instance made by the evaluated 3D CNN models, which allows a better understanding of the features learned. In this study, Grad-CAM++ [31] was employed for visualisation. This method provides the visual explanation of the model based on the pixel-wise weighting of the gradients of the convolution feature map. It measures the importance of each pixel in the convolutional feature map towards the final prediction of the model.

The classification score Y^c for class c can be expressed as:

$$Y^c = \sum_k w_k^c \sum_i \sum_j \sum_h A_{ijh}^k , \quad (7)$$

where A_{ijh}^k is the feature map of a particular spatial location (i, j, h), w_k^c is the weight for the feature map A^k and class c.

The class-based saliency map M^c used for the final visual explanation can be expressed as:

$$M_{ijh}^c = relu(\sum_k w_k^c A_{ijh}^k). \quad (8)$$

In the Grad-CAM++ [31], the weights w_k^c is calculated by a weighted average of the pixel-wise gradients, which can be written as:

$$w_k^c = \sum_i \sum_j \sum_h \alpha_{ijh}^{kc} \, relu\left(\frac{\partial Y^c}{\partial A_{ijh}^k}\right) , \quad (9)$$

where α_{ijh}^{kc} is the weighting co-efficient and the $\frac{\partial Y^c}{\partial A_{ijh}^k}$ is the pixel-wise gradient for feature map A^k and class c.

Considering Eq. (9), Eq. (7) can be rewritten as:

$$Y^c = \sum_k [\sum_a \sum_b \sum_d \alpha_{abd}^{kc} \, relu\left(\frac{\partial Y^c}{\partial A_{abd}^k}\right)] \sum_i \sum_j \sum_h A_{ijh}^k , \quad (10)$$

where (a, b, d) and (i, j, h) are iterators for the same activation map A^k for avoiding confusion. $relu$ has been dropped in the derivation since the function of which is as a threshold for allowing the gradients to flow back. Taking partial derivative A_{ijh}^k twice on both sides:

$$\frac{\partial^2 Y^c}{(\partial A_{ijh}^k)^2} = 2\alpha_{ijh}^{kc} \frac{\partial^2 Y^c}{(\partial A_{ijh}^k)^2} + \sum_a \sum_b \sum_d A_{abd}^k (\alpha_{ijh}^{kc} \frac{\partial^3 Y^c}{(\partial A_{ijh}^k)^3}). \quad (10)$$

Based on Eq. (10), α_{ijh}^{kc} can be calculated as:

$$\alpha_{ijh}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ijh}^k)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{ijh}^k\right)^2} + \sum_a \sum_b \sum_d (A_{abd}^k \frac{\partial^3 Y^c}{(\partial A_{ijh}^k)^3})}. \quad (11)$$

Considering Eq. (11), Eq. (9) can then be rewritten as:

$$w_k^c = \sum_i \sum_j \sum_h \frac{\frac{\partial^2 Y^c}{(\partial A_{ijh}^k)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{ijh}^k\right)^2} + \sum_a \sum_b \sum_d (A_{abd}^k \frac{\partial^3 Y^c}{(\partial A_{ijh}^k)^3})} relu\left(\frac{\partial Y^c}{\partial A_{ijh}^k}\right). \quad (12)$$

IV. Dataset

To evaluate the proposed method, this study produced a new dataset, which contains the driver's head and hand movement footages captured by 2 cameras during the experiment. There are 6 classes in this dataset, including 4 types of NDAs and 2 types of DAs. 14 participants (12 male and 2 female) were recruited for this experiment who are from 8 different counties. The participants' age is in the range from 23 to 35. They were required to hold a valid UK driving license. The videos were recorded in different weather and lighting conditions including sunny, cloudy, rainy and snowy.

Fig. 4. Location of the mounted cameras.

A. Experiment Design

The vehicle used in the experiment was an instrumented Land Rover Discovery 5. The car was modified to accommodate both automated driving and human driving. During the experiment, the vehicle is in automated driving mode and following a designed route in the enclosed roads. To ensure safety, a steering wheel and a set of pedals were added in the back seat of the vehicle, which allows the safety driver to intervene and override the autonomous system. The participants were required to engage in some activities while the vehicle is under the automated driving mode. After a period of time, the driver was asked to take over the vehicle and drive for 2 minutes. Four types of NDA investigated in this study are reading news, watching videos, playing games and answering questionnaires using a tablet. These activities were selected by considering the outcomes from surveys [33], [34]. The DAs considered in this study are road checking and driving. For each participant, the engagement of each activity (4 types of NDA and road checking) lasted 5 to 9 minutes followed by a 2 minutes driving process, which considered as one single trial. There are 5 trials per participant. The data of 4 NDA classes were extracted from the corresponding trials. The data for the road checking class contains the data extracted from the road checking trial and the data of the road checking behaviour during the NDA engagement trials. The data for driving was obtained by extracting the data where the participant was driving the vehicle after the take-over.

B. Camera Setup

The employed 2 cameras for monitoring the driver's behaviour in the experiment were Garmin Virb Action Camera, which provides the videos with 1920 × 1440 pixels spatial resolution and frames were sampled at 30 frames per second (fps). The front camera, facing the driver's face, is used to extract the driver's head movement and recognise whether the

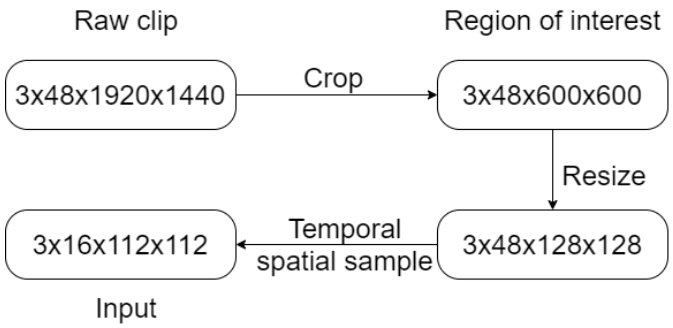


Fig.5. Data pre-process flowchart. The data format is presented as a four-dimensional tensor as $c \times l \times h \times w$, where c is the number of channels, l is the number of frames in the clip, h and w are the height and width of images, respectively.

driver is engaging with NDAs or DAs. The rear camera was mounted on the roof of the vehicle between two front seats to record the driver's hand movement. The location of the cameras is shown in Fig. 4. A flashing red LED light was employed for synchronisation, which can be seen in the view of both cameras.

C. Data Pre-processing

In the dataset for the driver activity recognition framework, a single instance, denoted by I, contains a pair of synchronised frame stacks (I_f, I_r) from the front camera and rear camera, respectively. The recorded video from each camera was split into several clips. We removed some bad clips which contains the participant's behaviour during the activity transition. The activity is difficult to be determined in such clips, such as the mixture of road-checking playing games, etc. As shown in Fig. 5, there are 48 frames in each clip, which were cropped with a 600×600 region of interest and further resized into 128×128. The dimension of the frames for each clip is $3 \times 48 \times 128 \times 128$. Then the 16 adjacent frames were randomly sampled and used as an input instance of I_f or I_r. The size of an input instance is $3 \times 16 \times 112 \times 112$. There are 7960 pairs of instances for 6 classes in total. The distribution of all these classes is answering questionnaires (1336), road checking (1320), driving (1268), playing games (1356), reading (1422) and watching videos (1258). The data were randomly split into 5 different segments for cross-validation based on participants. For each split, the data of 11 participants were used for training and the data of 3 participants were used for testing. The data distribution for 5 splits is split 1 (6158 for training and 1802 for testing), split 2 (6332 for training and 1628 for testing), split 3 (6176 for training and 1784 for testing), split 4 (6222 for training and 1738 for testing) and split 5 (6186 for training and 1774 for testing).

V. Train and Results

A. Training

The proposed method is compared with 3 state-of-the-art methods, including

(1) 3D ResNets (R3D) [18] that mainly utilises the basic $3 \times 3 \times 3$ residual block in the whole network to model the spatial-temporal information. Frequent usage of 3D convolution causes a higher computational cost.

(2) (2+1)D ResNets (R(2+1)D) [28] that factorises the 3D convolution of the residual block in R3D into two separate operations, which are a 2D spatial convolution and a 1D temporal convolution. Although such a structure doubles the number of nonlinearities to improve the model's capability of representing complex functions, the number of parameters and the computational cost is not decreased in comparison to the 3D CNN.

(3) Pseudo-3D ResNets (P3D) [27] that has the same method of factorisation with R(2+1D) but develops 3 blocks with different types of connection. It also adapts the bottleneck block in the network. However, the performance is not significantly improved than the simple and homogenous R(2+1D) network.

(4) The proposed DS3D ResNet.

For a fair comparison, all networks adapt 18 layers except P3D. Considering the specific design of the P3D architecture, the input size is $3 \times 16 \times 160 \times 160$. We also keep the same crop ratio from the raw frames as other models. The evaluated P3D model was built based on ResNets-50 architecture. All four models were trained from scratch on the same dataset. The size and computational complexity for these models are provided in Table I, which shows the proposed model has the lowest computational cost and smallest model size.

TABLE I
Comparison of the Model Size and the Computational Complexity. All Models Are Based on ResNes-18 Architecture.

Model	Parameters ($\times 10^6$)	FLOPs ($\times 10^9$)
R3D	33.1	83.1
R2+1D	33.2	85.2
The proposed DS3D	11.8	72.5

In the training process, Adam was used for parameter optimisation with the mini-batch size of 32. The initial learning rate was set as 0.001, which was divided by 10 after every 10 epochs. The whole training was completed in 35 epochs. The task of NDA engagement recognition adapts all the head movement dataset I_f. The tasks of NDA classification and DA classification use the corresponding data in the hand movement dataset I_r.

B. Results

The comparison results, based on the testing data for each spilt, are presented in Table II, which shows the models' accuracy for 3 tasks and the final fusion results. For the task of NDA engagement recognition (NDA or DA), the average accuracy of R3D for 5 splits is 87.49%. The performance of R2+1D and P3D is similar and around 90%. The proposed DS3D model achieves 93.74% average accuracy on this task. For the task of DAs classification (driving or road checking), all 3 state-of-the-art methods achieve similar performance while our model has at least 3% improvement than them. For the task of NDA classification (reading news, watching videos, playing games or answering questionnaires), the average accuracy of R3D, R2+1D and P3D models is 82.01%, 84.04% and 82.94%, respectively, while the accuracy of our model is 85.86%. For the final fusion result for the classification of all 6 activities, it can be observed that the proposed model achieves the best performance among the evaluated methods with at least 5% improvement.

TABLE II
ACCURACY OF THE EVALUATED MODELS ON THE PRODUCED DATASET.

Term	NDAs engagement recognition				DAs classification			
	R3D	**R2+1D**	**P3D**	**Our DS3D**	**R3D**	**R2+1D**	**P3D**	**Our DS3D**
Split 1	83.74%	87.79%	88.95%	93.90%	90.67%	93.08%	90.67%	95.71%
Split 2	87.78%	90.41%	89.07%	94.71%	91.70%	92.38%	92.21%	96.71%
Split 3	88.96%	90.92%	89.35%	93.57%	87.29%	90.28%	90.65%	90.46%
Split 4	88.15%	88.90%	92.28%	92.87%	91.36%	92.57%	89.46%	92.57%
Split 5	88.84%	90.19%	92.27%	93.63%	90.65%	86.06%	87.30%	93.47%
Average	87.49%	89.64%	90.38%	**93.74%**	90.33%	90.87%	90.06%	**93.78%**
Term	**NDAs classification**				**Fusion result**			
	R3D	**R2+1D**	**P3D**	**Our DS3D**	**R3D**	**R2+1D**	**P3D**	**Our DS3D**
Split 1	80.96%	83.81%	81.12%	87.20%	70.31%	74.64%	73.81%	83.46%
Split 2	84.19%	86.95%	84.95%	89.62%	75.43%	82.74%	77.27%	87.59%
Split 3	83.58%	84.38%	84.06%	85.10%	76.12%	78.98%	78.59%	82.90%
Split 4	81.10%	82.48%	82.14%	84.30%	74.51%	77.62%	76.41%	80.32%
Split 5	80.20%	82.60%	82.43%	83.10%	75.08%	76.16%	76.10%	82.47%
Average	82.01%	84.04%	82.94%	**85.86%**	74.29%	78.03%	76.44%	**83.35%**

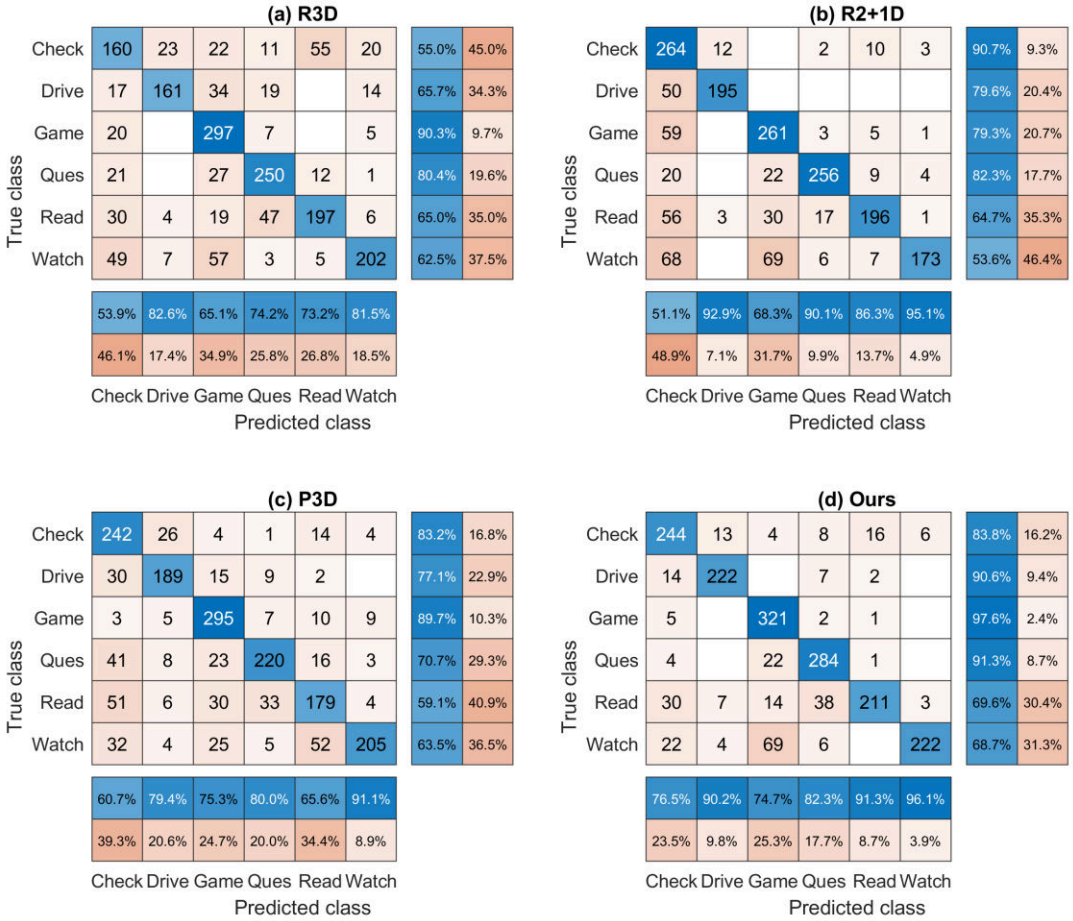


Fig. 6. Confusion matrix of the fusion results. The models used are trained on split 1. The precision and recall for each class are presented in the bottom and right of the figures, respectively. The classes presented in the figure refer to the activities named: road checking, driving, playing games, answering questionnaires, reading news and watching videos, successively.

The confusion matrices of the final fusion predictions are presented in Fig. 6. Precision and recall are used to evaluate the model in this study. Precision is the fraction of correct instances among the detected instances, while recall is the fraction of correctly detected instances [35]. For the category checking, the precisions of the 3 state-of-the-art models are around

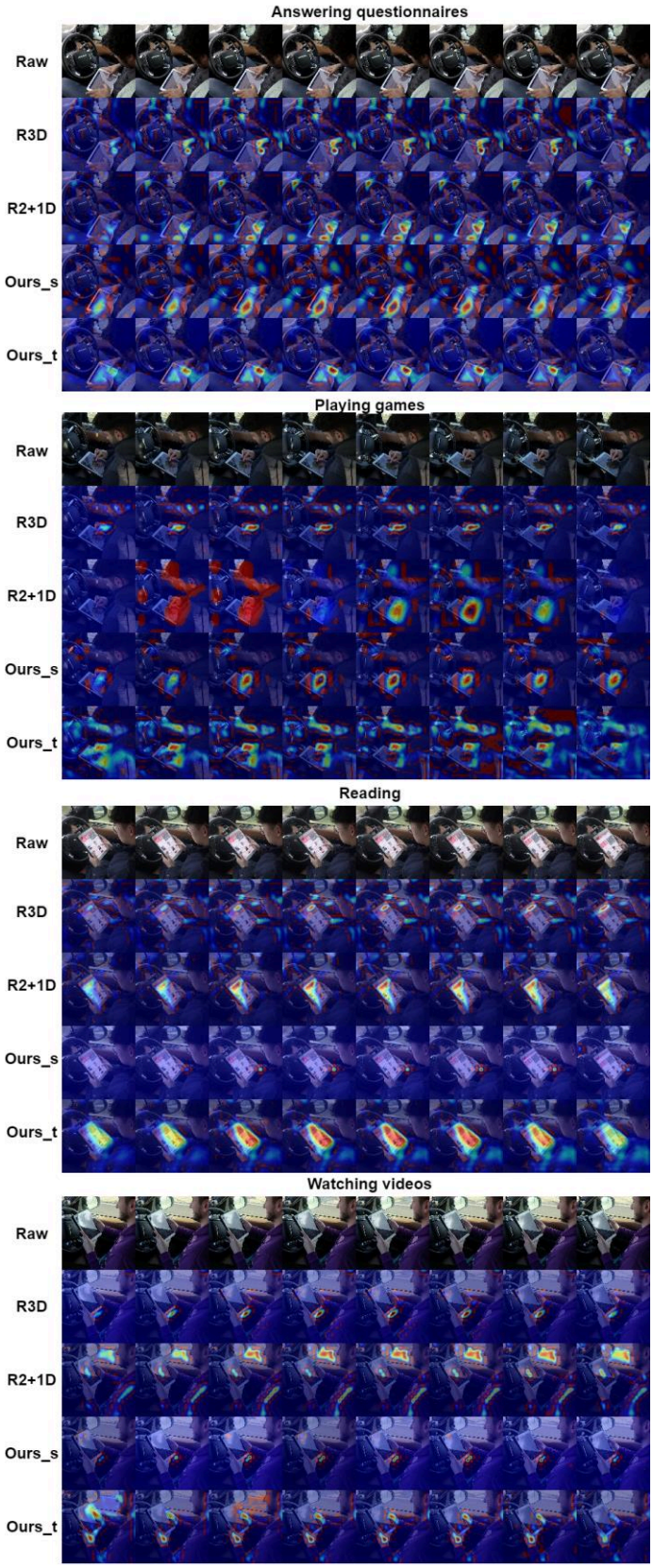


Fig. 7. Saliency maps of the prediction based on the last convolutional layer of Conv3 by using Grad-CAM++[28] for all NDAs. The first row of each activity is the raw frames imported into the network.

50%~60%. The main contribution of the false positive examples is from NDAs. It means that some NDAs have been predicted as DAs by being misclassified as checking, which suggests the poor performance of NDA engagement recognition for these models based on the participants head movement. For both DAs (checking and driving), the proposed DS3D achieves the best performance, specifically, 90.2% precision and 90.6% recall for driving. For NDA classification, answering questionnaires and playing games have a better performance than the other two activities for all 4 models. This is because, during these activities, it normally involves a high-frequency interaction between the participant's hand and the device. The superior performance of our model is benefited from the new structure design that enhances the spatial-temporal representations. The detailed contribution will be given in the next section with the saliency map. The recall of the other activities reading and watching videos for R3D, R2+1D, and P3D is around 60%~65%. The poor performance of these activities is due to similar observation associated with limited human-object interaction or hand movement in the temporal domain. The frames do not contain sufficient spatial-temporal information to make the right prediction for these activities. Even though, our model also outperforms the other evaluated models.

VI. Visualisation and Discussion

This section provides the visualisation results of the class-based saliency map in the hidden layer of the model trained on the dataset containing hand movement to explain the learned spatio-temporal feature. The images that contain facial information are not presented in this section due to the data protection policy.

In Fig. 7, the class-discriminative regions contributed from the hidden layer, Conv3, have been located, where the 16 frames are subsampled to 8 frames to save space. The regions in red correspond to a higher association for the class while the regions in blue represent weak relevance. It can be seen that the saliency regions have been highlighted on the frames based on the importance of the pixels. Specifically, the R3D model could learn the participant's hand movement when there is high-frequency interaction in the activity (answering questionnaires and playing games). For the activity like reading and watching movies, the learned features are mainly the edge of the object. The features used in the R2+1D model to predict are based on the context of the tablet. Both two models contain some noise such as steering wheel movement and background change of the side window. Comparing with these two models, the proposed DS3D model highlights the region of the hand movement concentratedly. The spatial stream of the proposed model (denoted as *Ours_s*) focuses on the short-time spatial feature learning. The temporal stream of the model (denoted as *Ours_t*) is to learn the small-region temporal feature. It can not only learn the short-time spatial feature, which is the high-frequency hand movement for the activities like answering questionnaires and playing games, but also the temporal feature, which is low-frequency interaction in the reading. Furthermore, it can give the right prediction based on the hand pose when there is a limited interaction during watching videos.

The saliency map results for DA engagement are presented in Fig. 8. For checking, the R3D model focuses on the edge of the steering wheel and the object. The R2+1 model highlights the region of the left hand. The spatial stream of our model

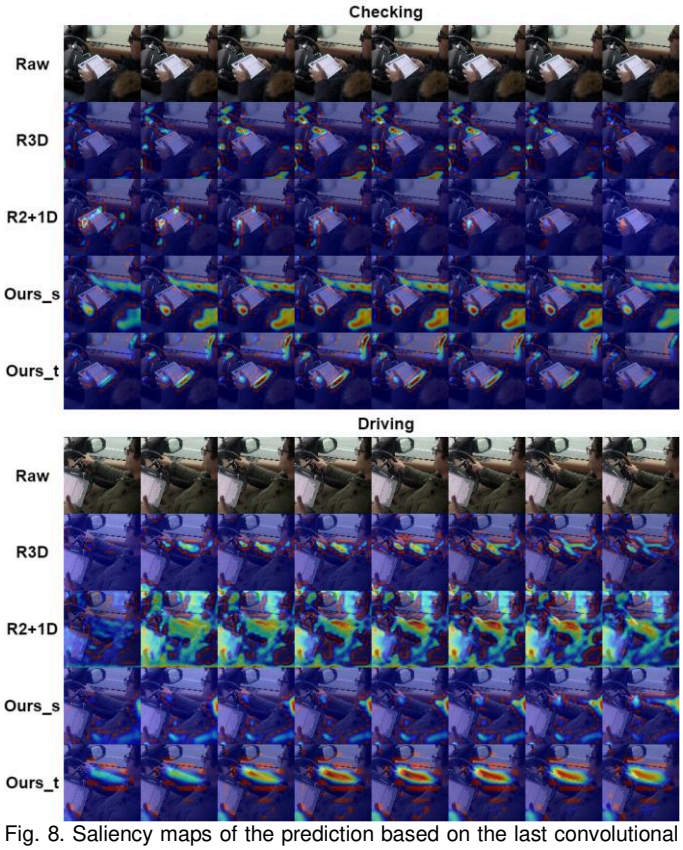


Fig. 8. Saliency maps of the prediction based on the last convolutional layer of Conv3 for all the DAs.

encodes the information of the hand pose and the door, while the temporal stream focuses on the edge of the head and the device. It explains the participant's road-checking behaviour during the NDA engagement where the participant headed up while holding the device on hands. In the driving category, the participant quickly steered the steering wheel with the right hand. The R3D model highlights the arm movement with its edge. The R2+1D model also learns the feature of the arm movement but with lots of noise. For the proposed model, the spatial stream captures the fast right-hand movement since it enhances the extraction of the short-time spatial change while the temporal stream mainly focuses on the participant's slight arm movement. From the perspective of model, the R3D model learns the semantically relevant features of the high-frequency interaction activity. But for the activities like reading, watching movies or road checking, the semanteme of feature is not clear. The R2+1D shows a better classification performance than R3D, however, the explainability of the learned feature is relatively weak. Collectively, it can be observed that, for all types of activity, the highlighted features learned by our model are more semantically relevant comparing with other models.

VII. Conclusion

In this paper, we propose a 2-feed 3D CNN based driver behaviour recognition system for the conditionally automated driving vehicle. Demonstrated by the testing results on the collected data, the introduced novel dual-stream 3D residual network (DS3D ResNet) presents a strong capability of encoding the spatial-temporal information for driver's behaviour. Specifically, the spatial stream extracts the short-time spatial features while the temporal stream focuses on learning the small-region temporal representation. This hypothesis has been successfully tested by visualising the saliency maps. Quantitative results demonstrate the superior performance of the proposed DS3D model against three state-of-the-art methods. From the perspective of NDA recognition, the activities with more human-object interaction can be classified more accurately due to the contained abundant spatial-temporal features. It should be noted that the evaluation was conducted on a novel driver activity dataset. Based on the visualisation results, we believe that the capability of the proposed DS3D model has not been fully explored using the current NDA dataset. The recognition of other NDAs with interaction in a higher frequency, for instance, phone typing, could benefit from this model. The application of the proposed method on a complehensive list of NDAs requires further study.

References

[1] C. Lv, X. Hu, A. Sangiovanni-Vincentelli, Y. Li, C. M. Martinez, and D. Cao, "Driving-Style-Based Codesign Optimization of an Automated Electric Vehicle: A Cyber-Physical System Approach," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 2965–2975, Apr. 2019.

[2] *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE International Standard J3016_201806, 2018.

[3] L. Yang et al., "A refined non-driving activity classification using a two-stream convolutional neural network," IEEE Sens. J., early access, Jun. 2020. doi: 10.1109/JSEN.2020.3005810.

[4] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4318–4327, Oct. 2020.

[5] J. Kim, H. S. Kim, W. Kim, and D. Yoon, "Take-over performance analysis depending on the drivers' non-driving secondary tasks in automated vehicles," *9th Int. Conf. Inf. Commun. Technol. Converg. ICT Converg. Powered by Smart Intell. ICTC 2018*, pp. 1364–1366, 2018.

[6] S. H. Yoon, Y. W. Kim, and Y. G. Ji, "The effects of takeover request modalities on highly automated car control transitions," *Accid. Anal. Prev.*, vol. 123, no. September 2017, pp. 150–158, 2019.

[7] K. Zeeb, A. Buchner, and M. Schrauf, "Is take-over time all that matters? the impact of visual-cognitive load on driver take-over quality after conditionally automated driving," *Accid. Anal. Prev.*, vol. 92, pp. 230–239, 2016.

[8] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, and F.-Y. Wang, "Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5379–5390, Jun. 2019.

[9] L. Yang, K. Dong, Y. Ding, J. Brighton, Z. Zhan, and Y. Zhao, "Recognition of visual-related non-driving activities using a dual-camera monitoring system," *Pattern Recognit.*, vol. 116, p. 107955, Aug. 2021.

[10] L. Meng *et al.*, "Interpretable Spatio-Temporal Attention for Video Action Recognition," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1513–1522.

[11] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, May 2020.

[12] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[13] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 4724–4733.

[14] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9692–9702, Dec. 2019.

[15] T. Huynh-The, C.-H. Hua, and D.-S. Kim, "Encoding Pose Features to Images With Data Augmentation for 3-D Action Recognition,"

IEEE Trans. Ind. Informatics, vol. 16, no. 5, pp. 3100–3111, May 2020.

[16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2556–2563.

[17] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," no. November, Dec. 2012.

[18] K. Hara, H. Kataoka, and Y. Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, vol. 2018-Janua, pp. 3154–3160.

[19] M. Martin, J. Popp, M. Anneken, M. Voit, and R. Stiefelhagen, "Body Pose and Context Information for Driver Secondary Task Detection," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, vol. 2018-June, no. Iv, pp. 2015–2021.

[20] Y. Xing *et al.*, "Identification and Analysis of Driver Postures for In-Vehicle Driving Activities and Secondary Tasks Recognition," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 1, pp. 95–108, Mar. 2018.

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[22] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Biochem. Pharmacol.*, vol. 32, no. 5, pp. 849–855, Jun. 2014.

[23] H. Xu, A. Das, and K. Saenko, "R-C3D: Region Convolutional 3D Network for Temporal Activity Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2319–2332, Mar. 2017.

[24] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, "Driver Distraction Identification with an Ensemble of Convolutional Neural Networks," *J. Adv. Transp.*, vol. 2019, Jan. 2019.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, vol. 2015 Inter, pp. 4489–4497.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 770–778.

[27] Z. Qiu, T. Yao, and T. Mei, "Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 2017-Octob, pp. 5534–5542.

[28] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[29] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, Feb. 2015.

[30] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Work. Track Proc.*, pp. 1–14, Dec. 2014.

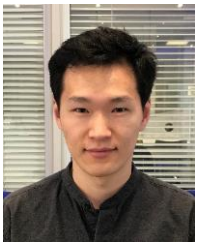
[31] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, vol. 2018-Janua, pp. 839–847.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[33] M. Sivak, B. Schoettle, "Motion Sickness in Self-Driving Vehicles," Transportation Res. Inst., Ann Arbor, Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep. UMTRI-2015-12, Apr. 2015.

[34] F. Naujoks, D. Befelein, K. Wiedemann, and A. Neukum, "A Review of Non-driving-related Tasks Used in Studies on Automated Driving," in *Advances in Intelligent Systems and Computing*, vol. 597, 2018, pp. 525–537.

[35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

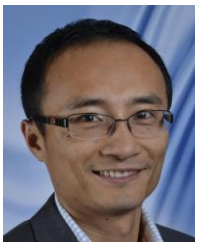


Lichao Yang was born in Shanxi, China. He received the M.Sc. degree in automotive mechatronics from Cranfield University, Cranfield, U.K., in 2018. He is currently working toward the Ph.D. degree in driver non-driving activities analysis at Through-Life Engineering Services Centre, Cranfield University, Cranfield, U.K.

Xiaocai Shan has been working toward a Ph.D. degree in the Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China. She is currently a visiting Ph.D. student at Through-Life Engineering Services Centre, Cranfield University, Cranfield, U.K. Her research interests are seismic signal processing and brain connectivity analysis of EEG data.

Chen Lv is currently an Assistant Professor at School of Mechanical and Aerospace Engineering, and the Cluster Director in Future Mobility Solutions at ERI@N, Nanyang Technology University, Singapore. He received the Ph.D. degree at the Department of Automotive Engineering, Tsinghua University, China in 2016. He was a joint PhD researcher at EECS Dept., University of California, Berkeley, USA during 2014-2015, and worked as a Research Fellow at Advanced Vehicle Engineering Center, Cranfield University, UK during 2016-2018.

James Brighton has over 22 years' experience relating to off road vehicle dynamics, terra-mechanics, tyre and track system modelling, advanced vehicle instrumentation and lightweight material structures and his current clients span the globe. His team is able to offer a wide range of vehicle related technical solutions from fundamental research through product design and prototype vehicle sub-system manufacture, supply, evaluation and testing across a wide range of industry sectors.



Yifan Zhao was born in Zhejiang, China. He received the PhD degree in Automatic Control and System Engineering from the University of Sheffield, UK in 2007.

He currently is a Senior Lecturer in Data Science at Cranfield University. His research interests are computer vision for automated vehicles, human behaviour analysis, super resolution, active thermography and nonlinear system identification.

Cranfield University

CERES https://dspace.lib.cranfield.ac.uk

School of Aerospace, Transport and Manufacturing (SATM) Staff publications (SATM)

2021-07-28

Learning spatio-temporal representations with a dual-stream 3-D residual network for nondriving activity recognition

Yang, Lichao

IEEE

Yang L, Shan X, Lv C, et al., (2022) Learning spatio-temporal representations with a dual-stream 3-D residual network for nondriving activity recognition. IEEE Transactions on Industrial Electronics, Volume 69, Number 7, July 2022, pp. 7405-7414

https://doi.org/10.1109/TIE.2021.3099254

Downloaded from Cranfield Library Services E-Repository